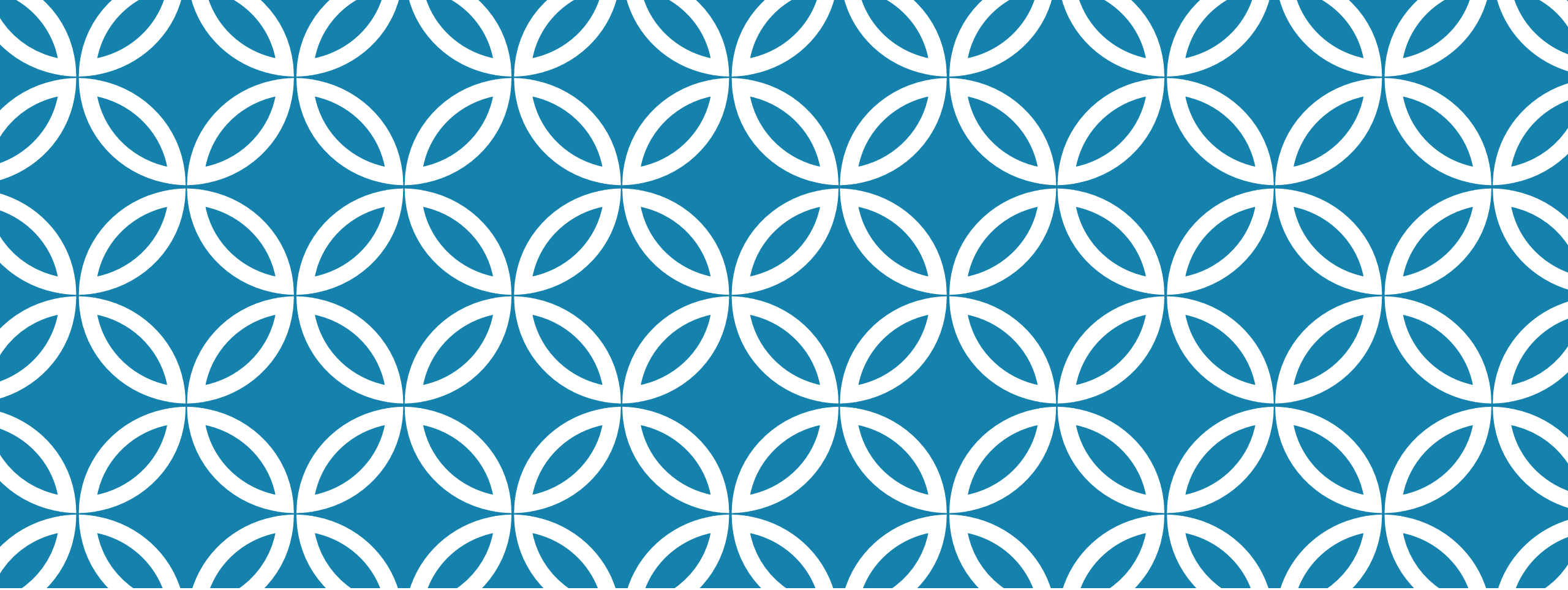


LEAD SCORING CASE STUDY

Prediction of Hot Leads for an Education Company X

Team Members: Arun, Prasad, Sai

CONTENTS

1. Problem Statement & Objective of the Study
2. Analysis Approach
3. Data Cleaning and EDA
4. Data Preparation
5. Model Building (RFE & Manual fine tuning) and Model Evaluation
6. Recommendations

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets

leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted, while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads. Such that, the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Goals of the Case Study

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Objectives:

To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

The CEO has given a target lead conversion rate to be around 80%.

ANALYSIS APPROACH

1. Data Cleaning and EDA
2. Data Preparation
3. Model Building and Evaluation
4. Predictions on Test Data
5. Recommendations

DATA CLEANING AND EDA

"Select" level represents null values for some categorical variables.

Columns with over 40% null values were dropped. Missing values in categorical columns were handled based on value counts and certain considerations. Drop columns that don't add any insight or value to the study objective. Imputation was used for some categorical variables.

Additional categories were created for some variables. Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.

Numerical data was imputed with mode after checking distribution.

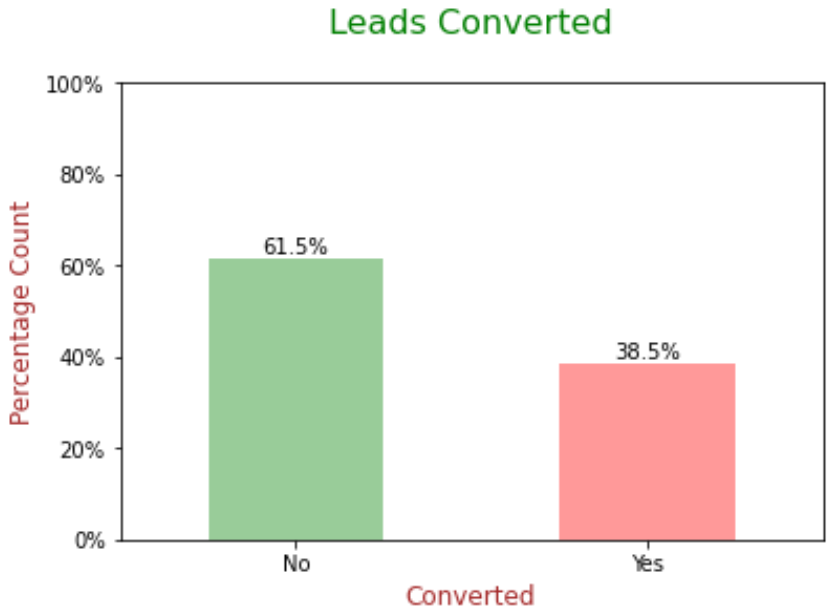
Outliers in TotalVisits and Page Views Per Visit were treated and capped.

Invalid values were fixed and data was standardized in some columns, such as lead source. Low frequency values were grouped together to “Others”. Binary categorical variables were mapped.

Other cleaning activities were performed to ensure data quality and accuracy.

Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc.

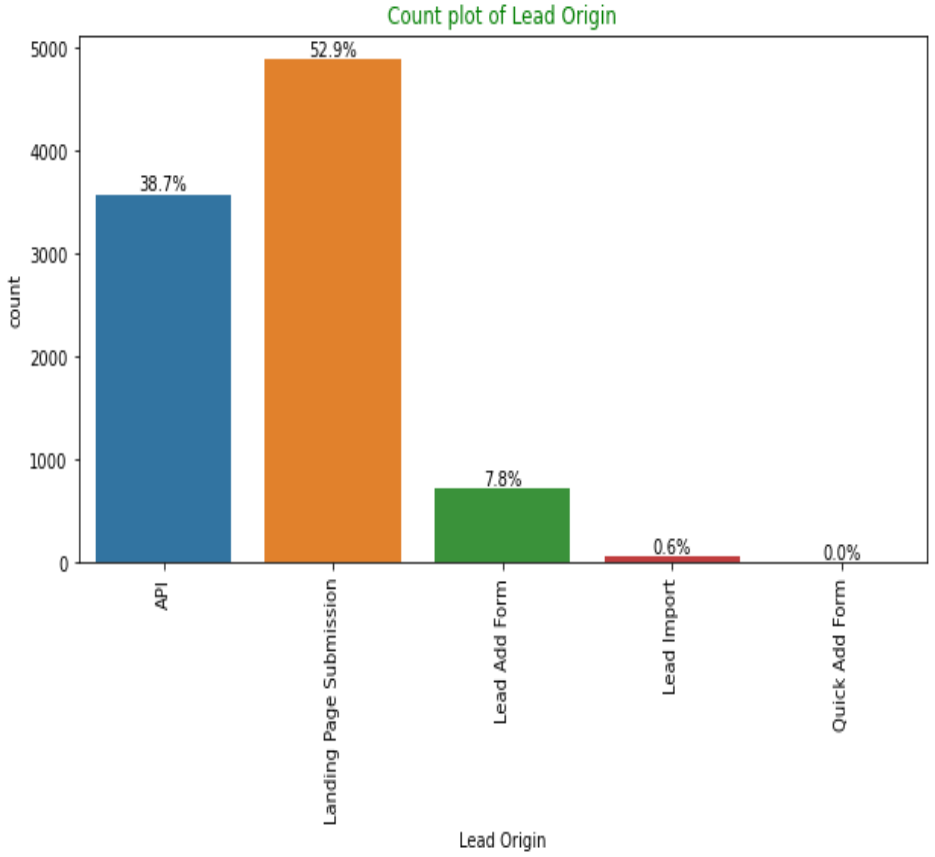
EDA

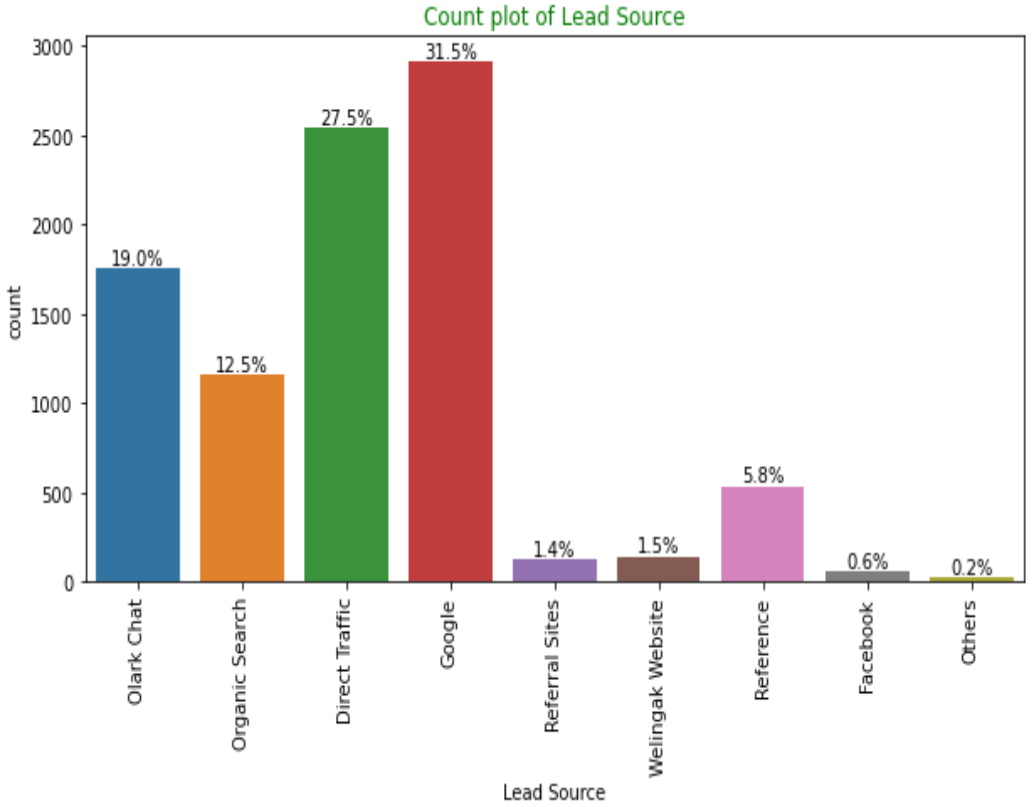


38.5% of the people have converted to leads while 61.5% of the people didn't convert to leads.

EDA

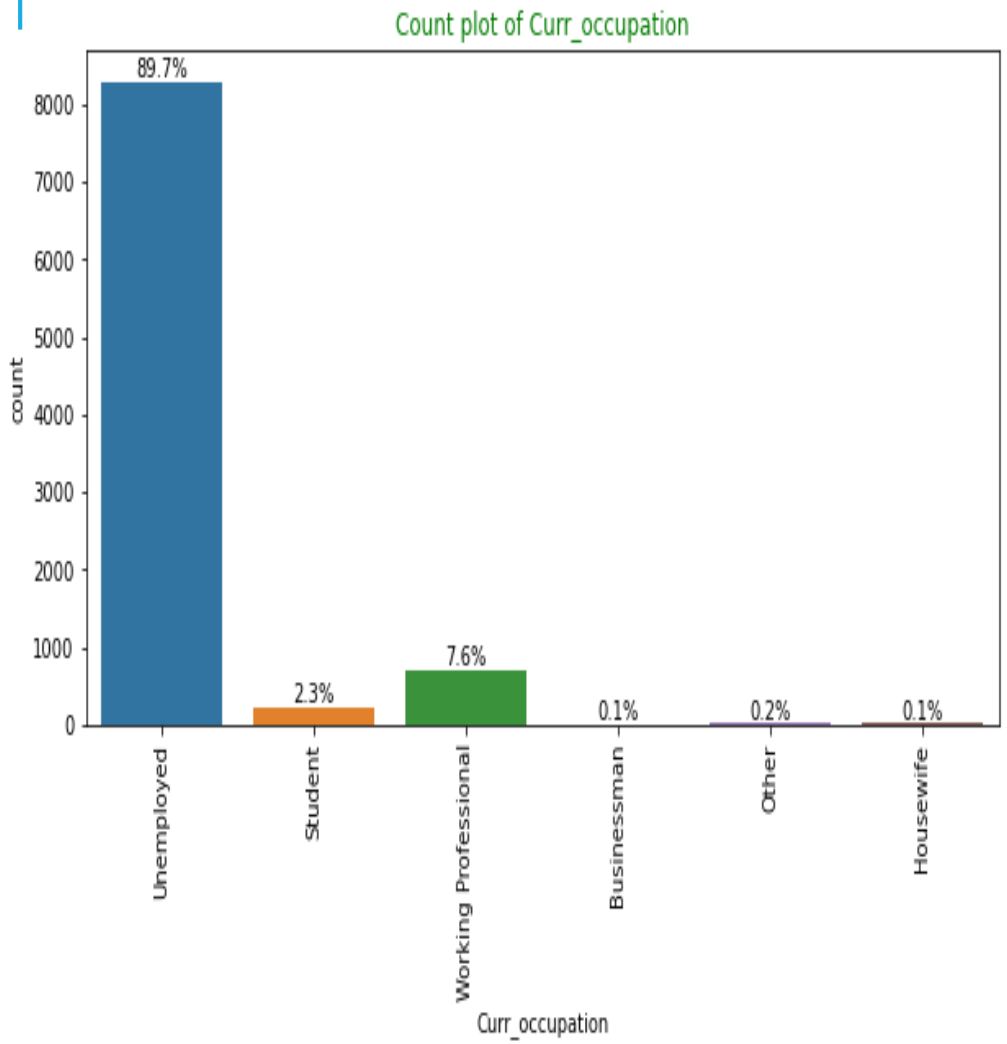
Univariate Analysis – Categorical Variables:

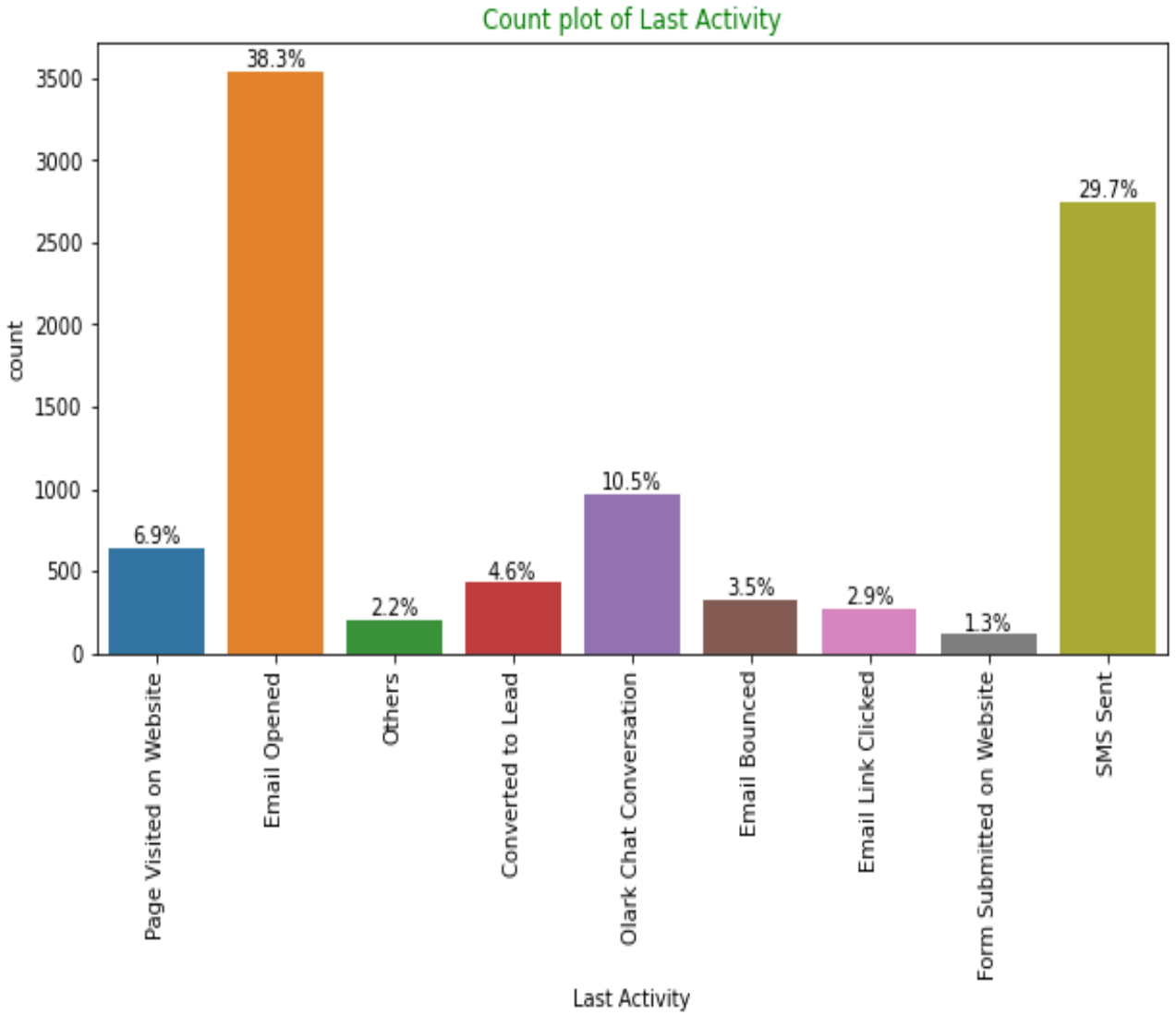




EDA

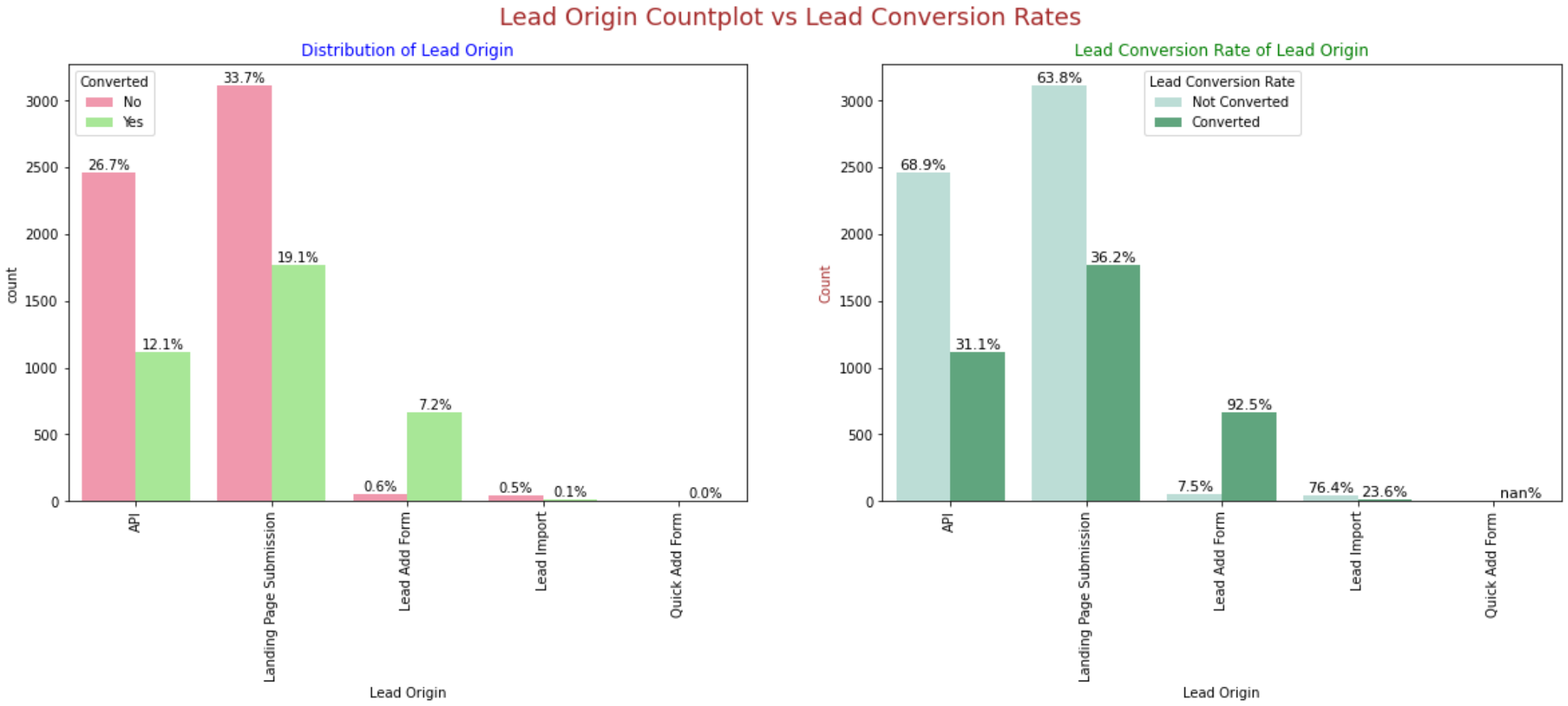
Univariate Analysis – Categorical Variables:





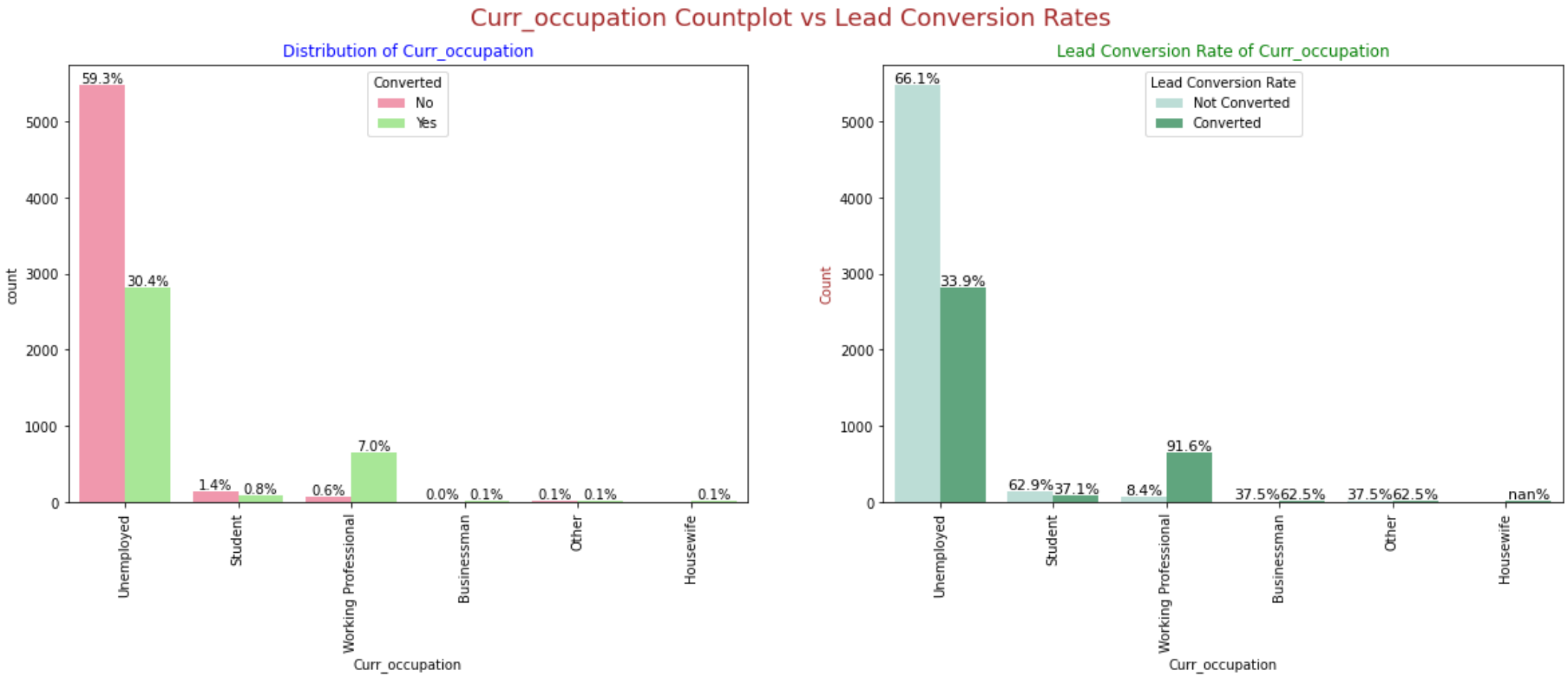
EDA

Bivariate Analysis for Categorical Variables



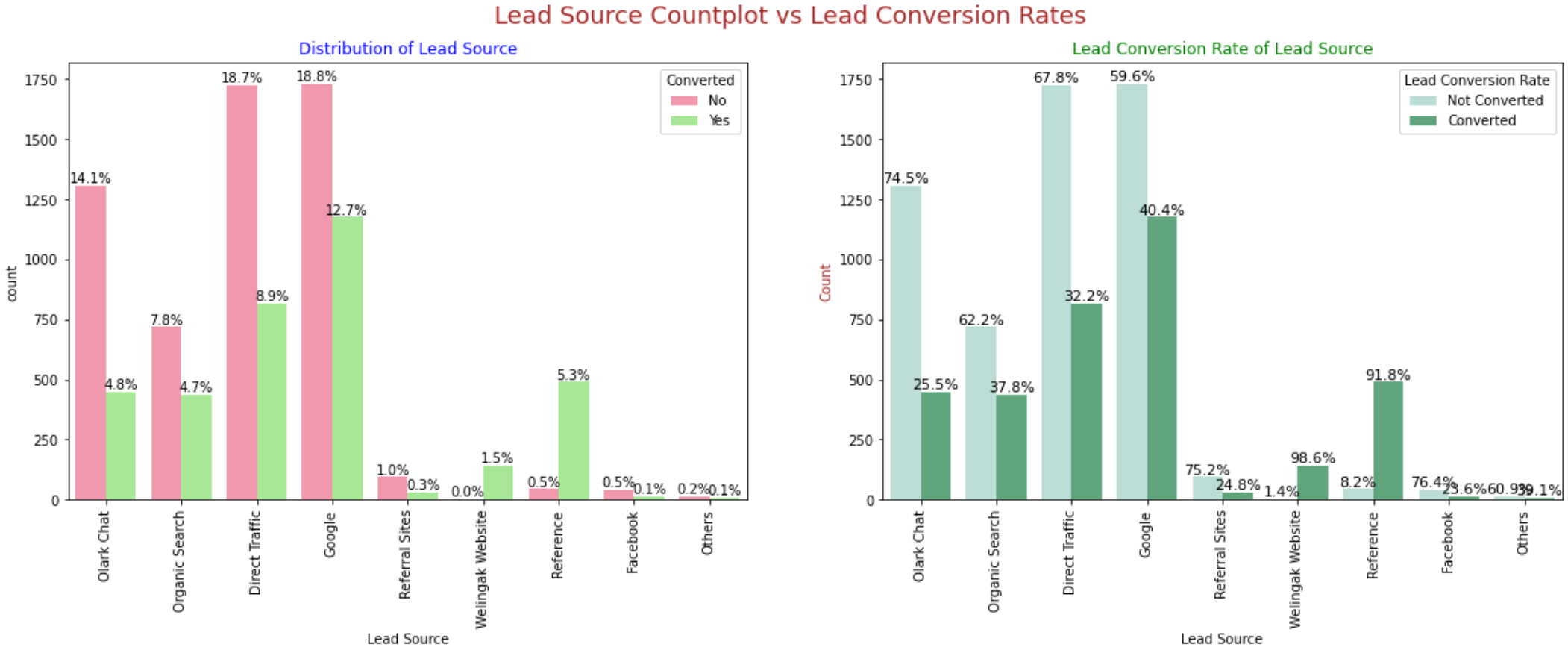
EDA

Bivariate Analysis for Categorical Variables



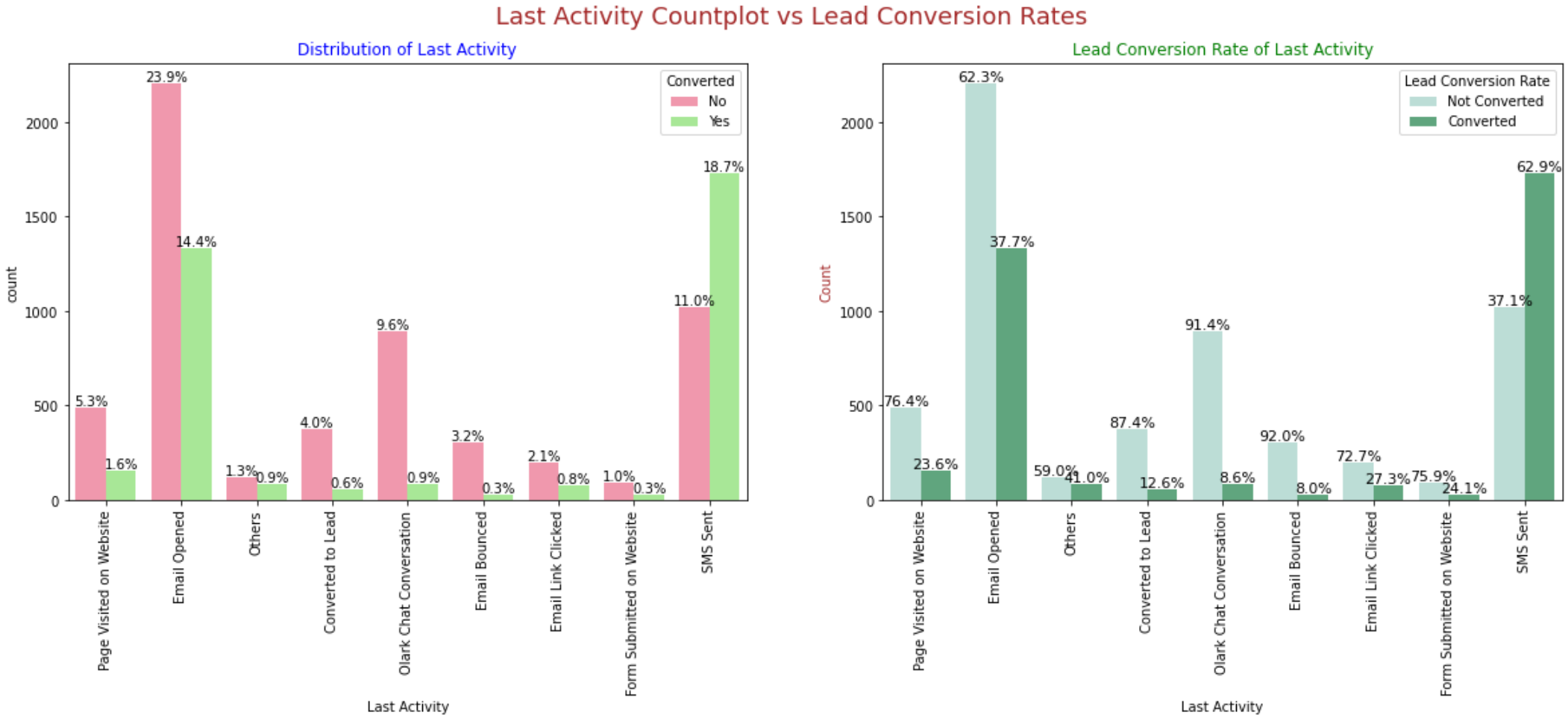
EDA

Bivariate Analysis for Categorical Variables



EDA

Bivariate Analysis for Categorical Variables



DATA PREPARATION

Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation

Splitting Train & Test Sets ○ 70:30 % ratio was chosen for the split

Feature scaling ○ Standardization method was used to scale the features

Checking the correlations ○ Predictor variables which were highly correlated with each other were dropped.

MODEL BUILDING AND EVALUATION

Used RFE to perform variable selection.

Built a Logistic Regression Model.

Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.

Final Model 4 was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5. logm4 was selected as final model with 13 variables and used it for making prediction on train and test set.

Confusion matrix was made and cut off point of 0.34 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less performance metrics around 75%. Lead score was assigned to train data using 0.34 as cut off.

RECOMMENDATIONS:

We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.

Lead Source_Welingak Website	5.388662
Lead Source_Reference	2.925326
Curr_occupation_Working Professional	2.669665
Last Activity_SMS Sent	2.051879
Last Activity_Others	1.253061
Total Time Spent on Website	1.049789
Last Activity_Email Opened	0.942099
Lead Source_Olark Chat	0.907184

RECOMMENDATIONS:

Strategies to be developed to attract high-quality leads from top-performing lead sources.

Optimize communication channels based on lead engagement impact.

More budget can be spent on Welingak Website in terms of advertising, etc.

Incentives/discounts for providing references that convert to leads.

Working professionals to be targeted as they have high lead conversion rate.